BLAST, Profile, and PSI-BLAST

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

2006

Problems of Using Dynamic Programming to Search Large Sequence Database

- Search homologs in DNA and protein database is often the first step of a bioinformatics study.
- Local DP is too slow for large sequence database search such as Genbank and SwissProt.

  Each DP search can take hours.
- Most DP search time is wasted on unrelated sequences or dissimilar regions.
- Developing fast, heuristic sequence comparison methods for database search is important.

Fast Sequence Search Methods

- All successful, rapid sequence comparison methods are based on a simple fact: similar sequences /regions **share some common words**. (This can improve sequence database search VERY significantly. Why?)
- First such method is FASTP (Pearson & Lipman, 1985)
- Most widely used methods are BLAST (Altschul et al., 1990) and PSI-BLAST (Altschul et al., 1997).

Basic Local Alignment Search Tool

(S. Altschul, W. Gish, W. Miller, E. Meyer and D. Lipman)

1. Compile a list of words for a query
2. Scan sequences in database for hits
3. Extending hits

David Lipman

Stephen Altschul

Step 1: Compile Word List

- Words: w-mer with length w.
- Protein 4-mer and DNA 12-mer

Query:

DSRSKGEPRDSGTLQSQEAKAVKKTSLFE

Words: DSRS, SRSK, RSKG, KGEP….

Notes: For DNA, use exact words appearing in the query. For protein, also include words similar to the words in the query (score > T =14)

Step 2: Scan database

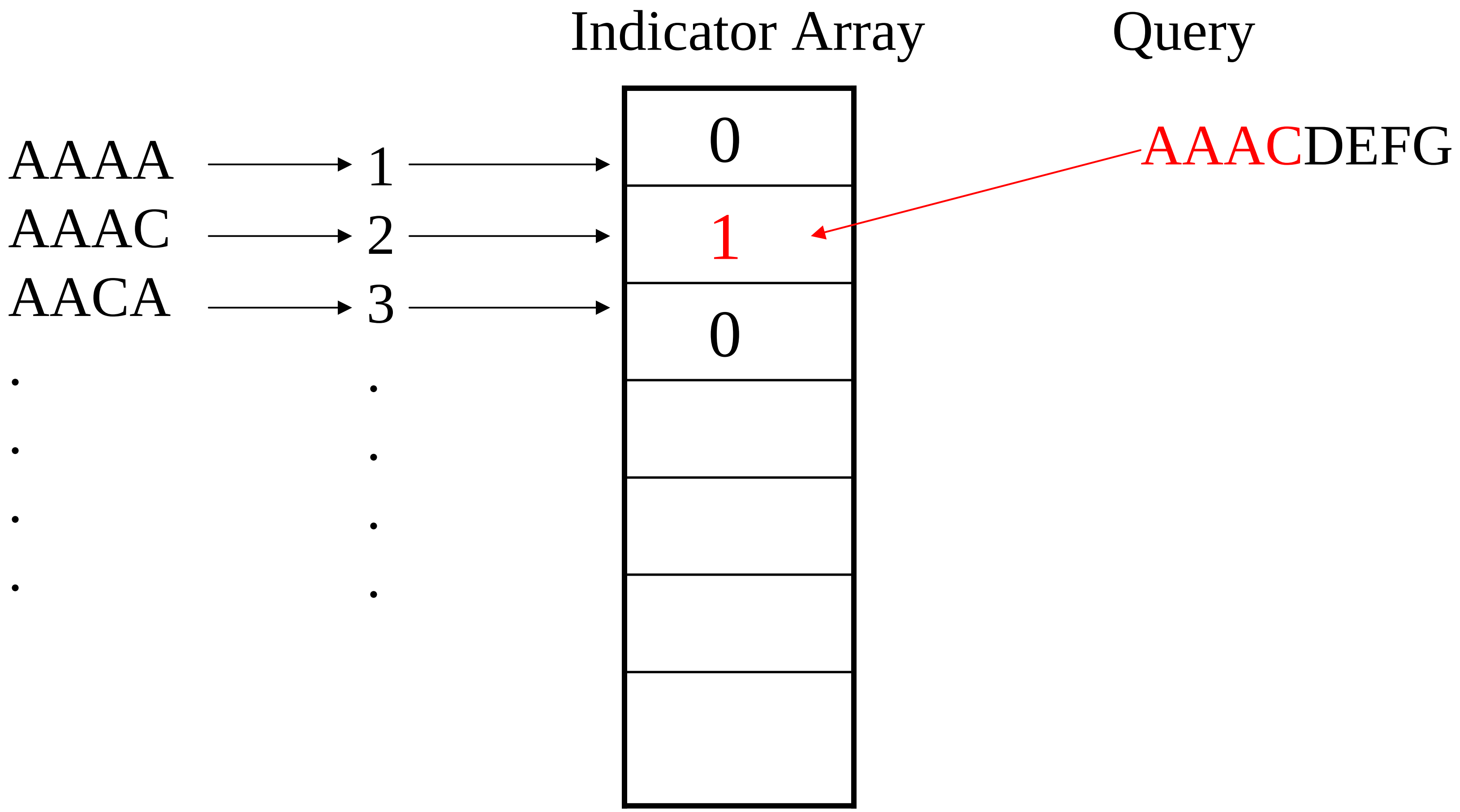
Classic problem: find occurrence of a list of words in a sequence.

- Integer indexing approach (hashing)

- Deterministic finite automaton or finite state machine. (faster)

Integer Indexing Approach

- Total number of protein 4-mer is: 20 * 20 * 20 * 20 = 20^4 = 160000
- Assign each 4-mer to an integer index in [1,160000]
- Create an array of 16000 elements representing 16000 possible words. Only words in query has value 1. Others 0.



Indicator Array

Query

AAAA → 1 → 0

AAAC → 2 → 1

AACA → 3 → 0

AAACDEFG

For each word in a sequence in database, convert it to an index, and then check indicator array to see if the element is 1, which indicates a word match.

Step 3: Extension

- Extend words on both ends
- Terminate the process when we reach a segment pair whose score falls a certain distance below the best score found for shorter extensions.
- Depart from the ideal of finding a guaranteed Maximum Segment Pair, but the added inaccuracy is negligible.
- Report significant MSP according to extreme value distribution

Example of extension

Query: DSRSKGEPRDSGTLQSQEAKAVKKTSLFE

Words: DSRS, **SRSK**, RSKG, KGEP….

Database Sequence: PE**SRSK**GEPRDSGKKQMDSOKPD

Maximum Segment Pair: **ESRSKGEPRDSG**

Gapped Extension and Performance

- Use dynamic programming to extend hits so as to allow gaps in the resulting alignments.
- On average, 40 time faster than DP for two sequences. But in reality, it is much faster for database search.
- Comparable sensitivity
- Fewer false positives

Why is BLAST so successful?

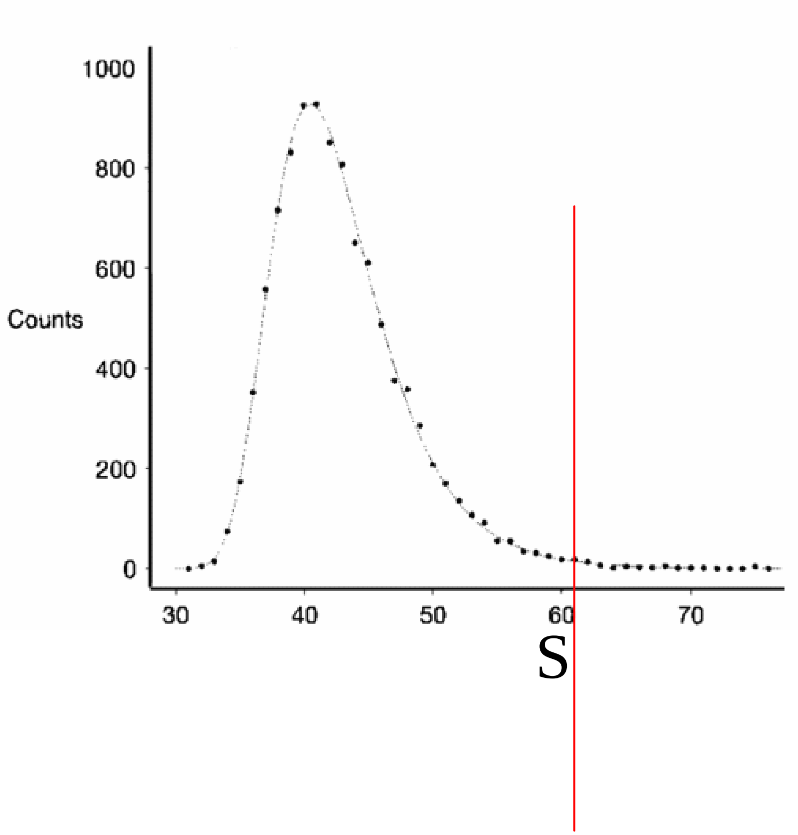
- Address a fundamental problem
- Simple, yet powerful idea
- Well founded in theory (words-string matching, hashing, random process of Maximum Segment Pair)
- Implementation tricks -> super speed
- Sacrifice a little accuracy for speed practically (good heuristics)

Usage of BLAST

- Versions: BLASTP, BLASTN, BLASTX (translated)
- Sequence Databases: NR, PDB, SwissProt, Gene databases of organisms, or your own databases
- Input format: FASTA
- Expectation value
- Low complexity
- Similarity matrix (PAM or BLOSUM)
- Output format

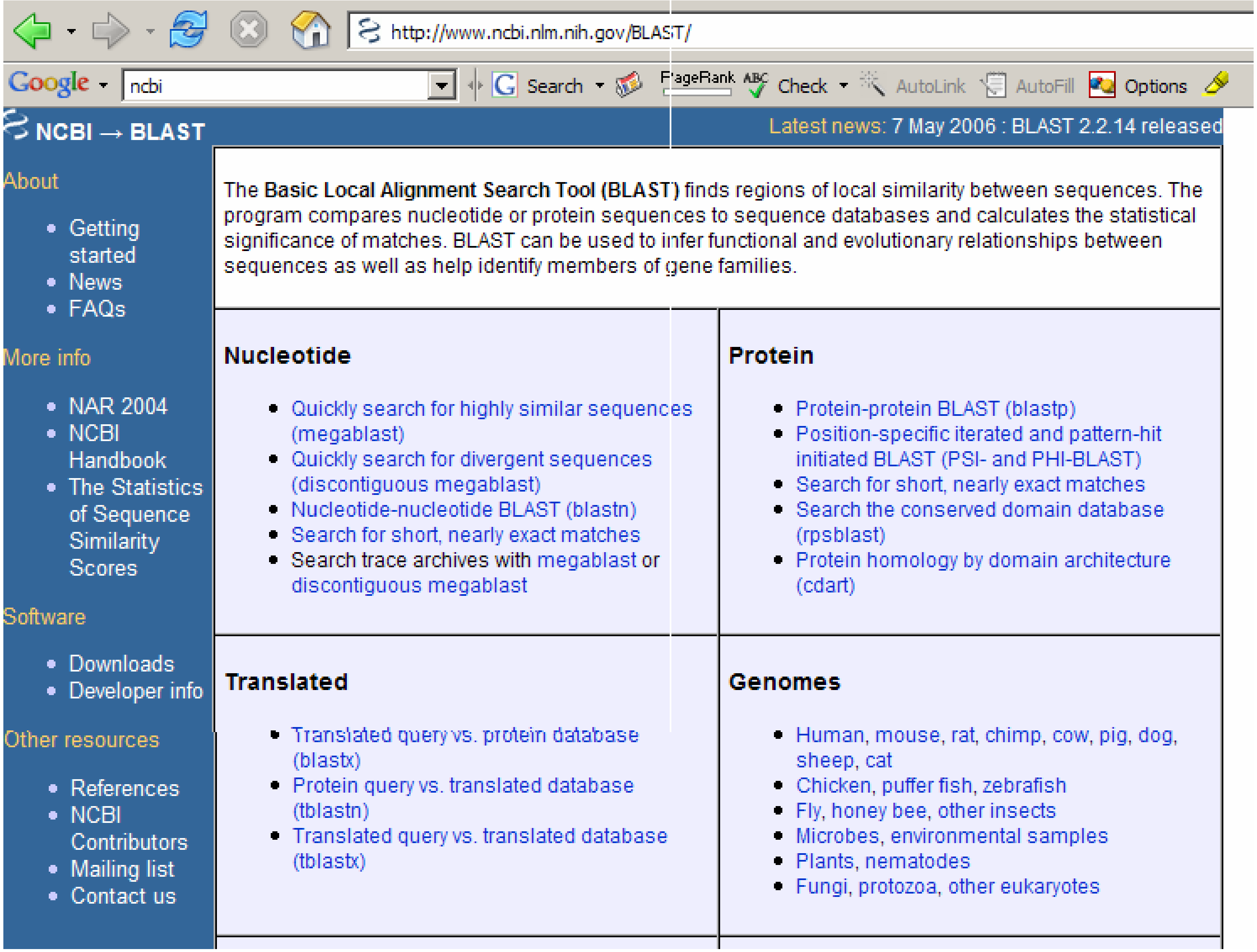
Input Format and E-Value

- P-value
- E-value = database size (n) * p-value
- Common threshold: 0.01



P-value = Prob(score >=S)

NCBI Online Blast



NCBI → BLAST

Latest news: 7 May 2006 : BLAST 2.2.14 released

About

- Getting started
- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

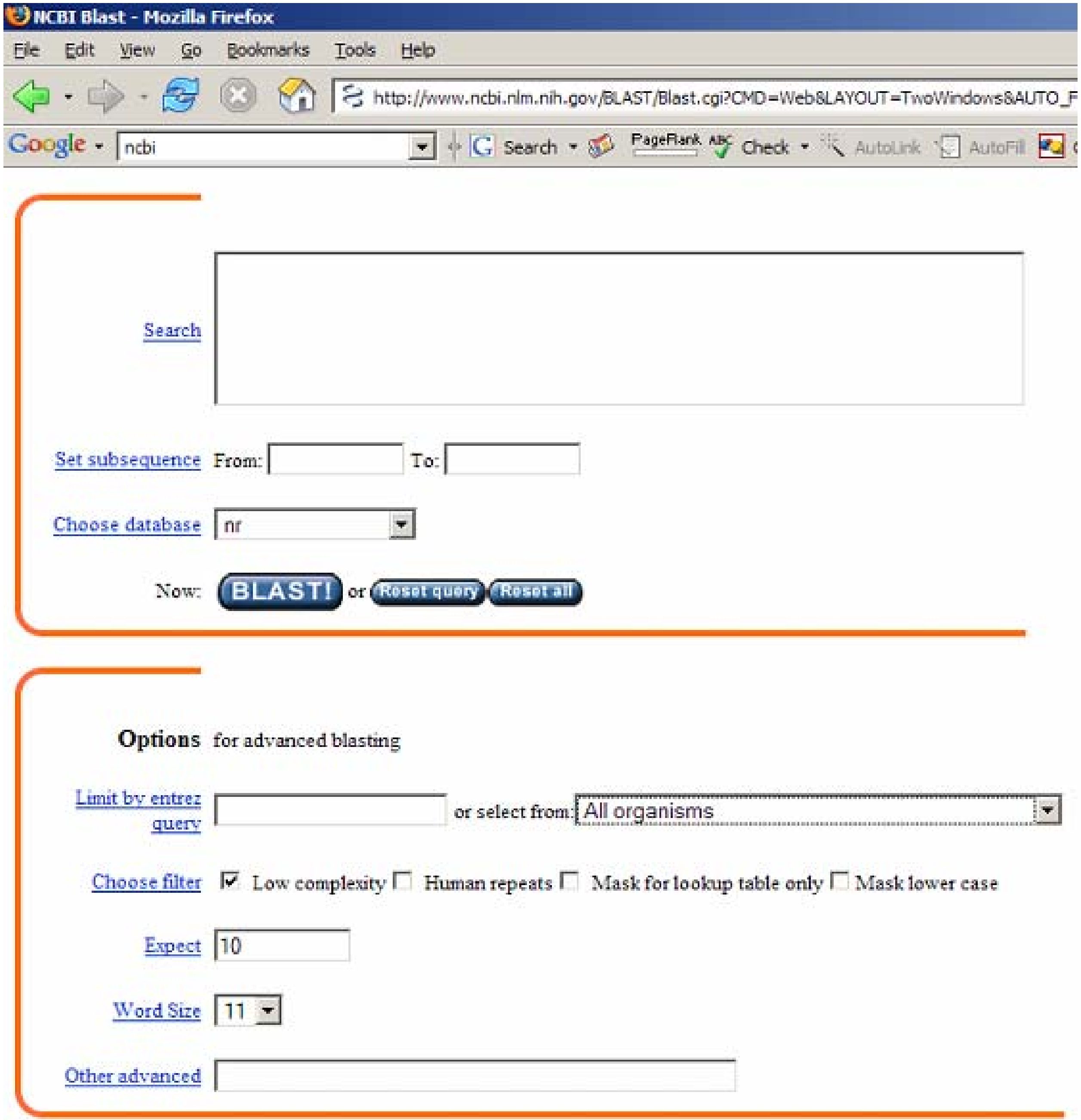
Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

DNA Blast



http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_F

ncbi

Search

Set subsequence From: To:

Choose database nr

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

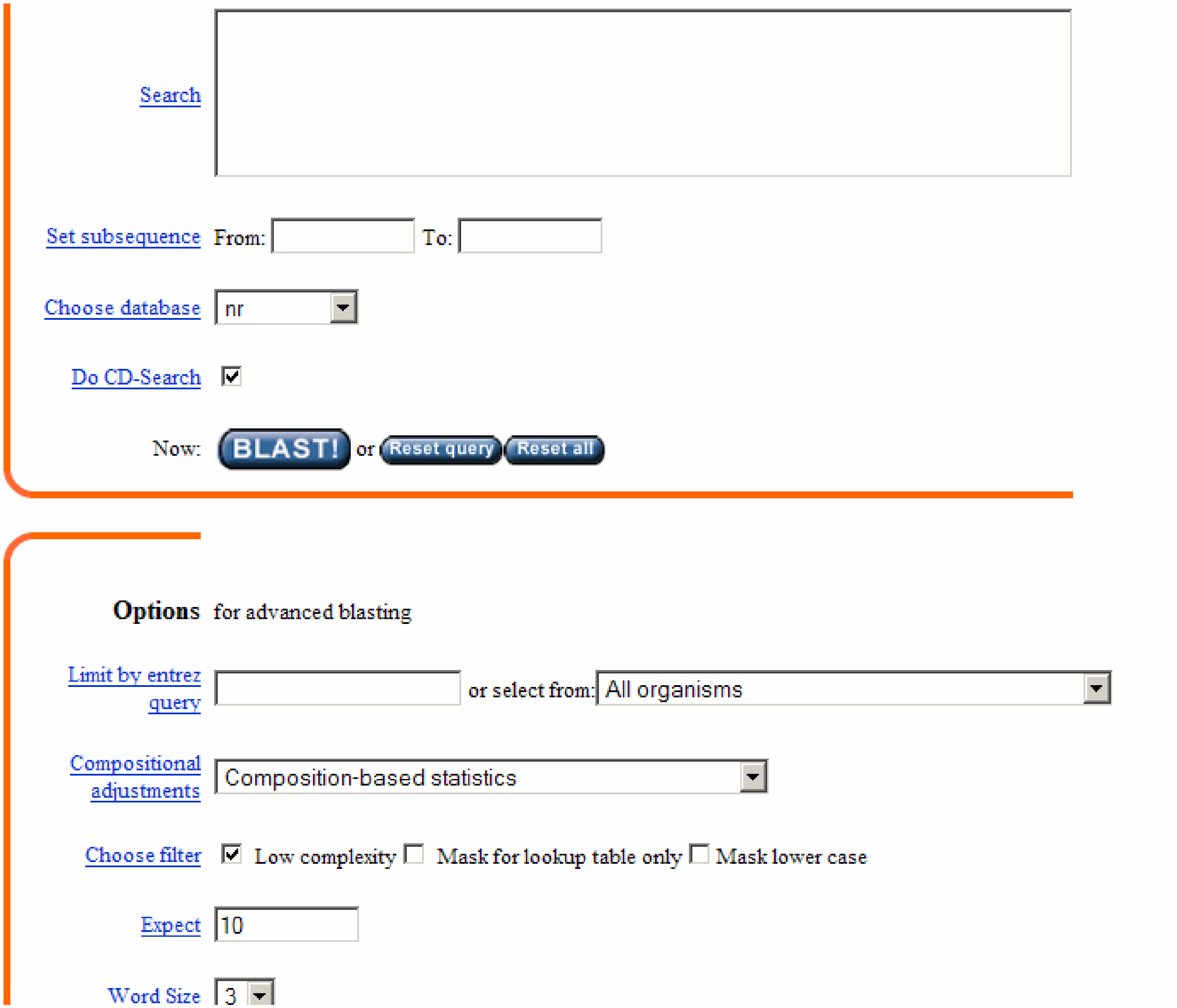
Choose filter ☑ Low complexity □ Human repeats □ Mask for lookup table only □ Mask lower case

Expect 10

Word Size 11

Other advanced

Protein Blast



Search

Set subsequence From: To:

Choose database: nr

Do CD-Search ☑

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query; or select from: All organisms

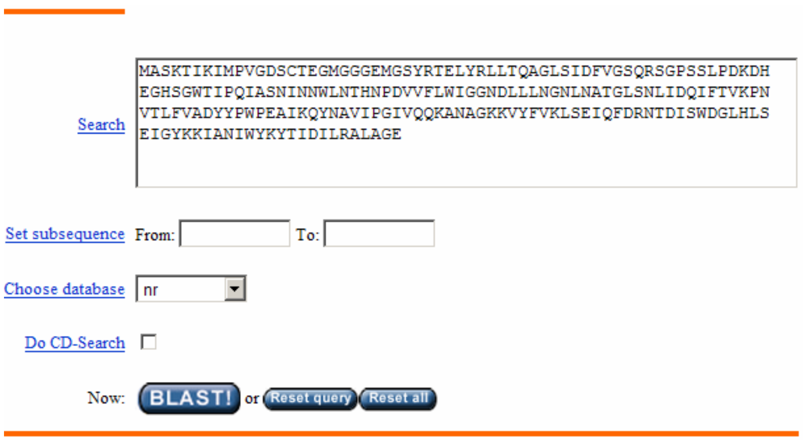
Compositional adjustments: Composition-based statistics

Choose filter ☑ Low complexity □ Mask for lookup table only □ Mask lower case

Expect: 10

Word Size: 3

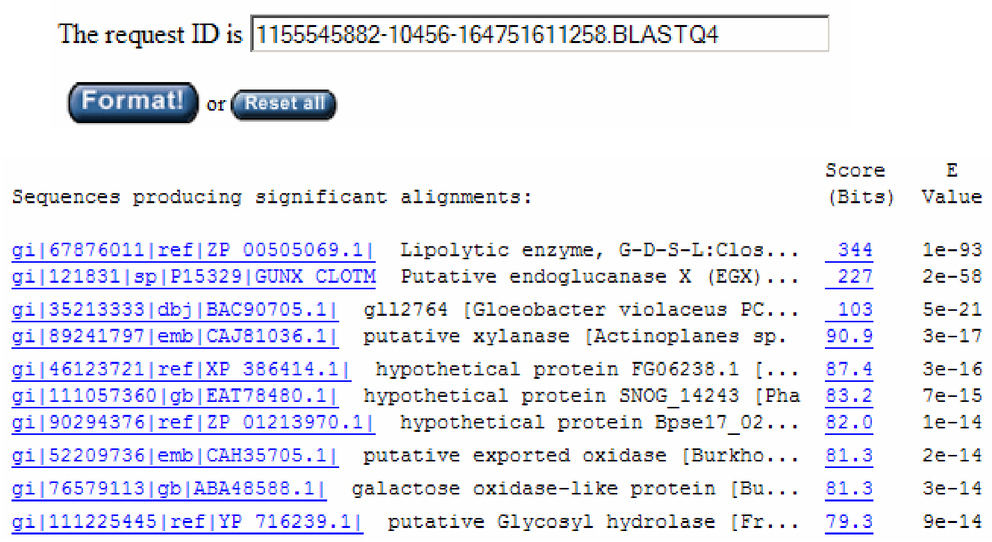
Output Format



```
MASKTIKIMPVGDSCTEGMGGGEMGSYRTELYRLLTQAGLSIDFVGSQRSGPSSLPDKDH
EGHSGWTIPQIASNINNWLNTHNPDVVFLWIGGNDLLLNGNLNATGLSNLIDQIFTVKPN
VTLFVADYYPWPEAIKQYNAVIPGIVQQKANAGKKVYFVKLSEIQFDRNTDISWDGLHLS
EIGYKKIANIWYKYTIDILRALAGE
```

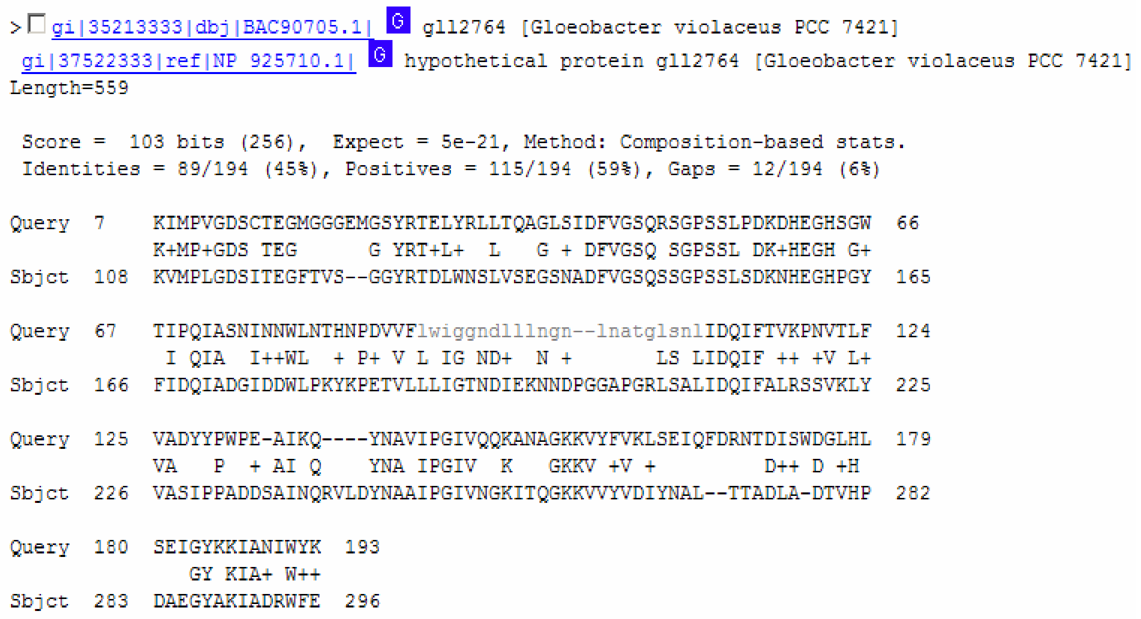
nr

□


The request ID is 1155545882-10456-164751611258.BLASTQ4

```
                                                              Score    E
Sequences producing significant alignments:                   (Bits)  Value

gi|67876011|ref|ZP_00505069.1|  Lipolytic enzyme, G-D-S-L:Clos...   344   1e-93
gi|121831|sp|P15329|GUNX_CLOTM  Putative endoglucanase X (EGX)...   227   2e-58
gi|35213333|dbj|BAC90705.1|  gll2764 [Gloeobacter violaceus PC...   103   5e-21
gi|89241797|emb|CAJ81036.1|  putative xylanase [Actinoplanes sp.   90.9   3e-17
gi|46123721|ref|XP_386414.1|  hypothetical protein FG06238.1 [...  87.4   3e-16
gi|111057360|gb|EAT78480.1|  hypothetical protein SNOG_14243 [Pha  83.2   7e-15
gi|90294376|ref|ZP_01213970.1|  hypothetical protein Bpse17_02...  82.0   1e-14
gi|52209736|emb|CAH35705.1|  putative exported oxidase [Burkho...  81.3   2e-14
gi|76579113|gb|ABA48588.1|  galactose oxidase-like protein [Bu...  81.3   3e-14
gi|111225445|ref|YP_716239.1|  putative Glycosyl hydrolase [Fr...  79.3   9e-14
```

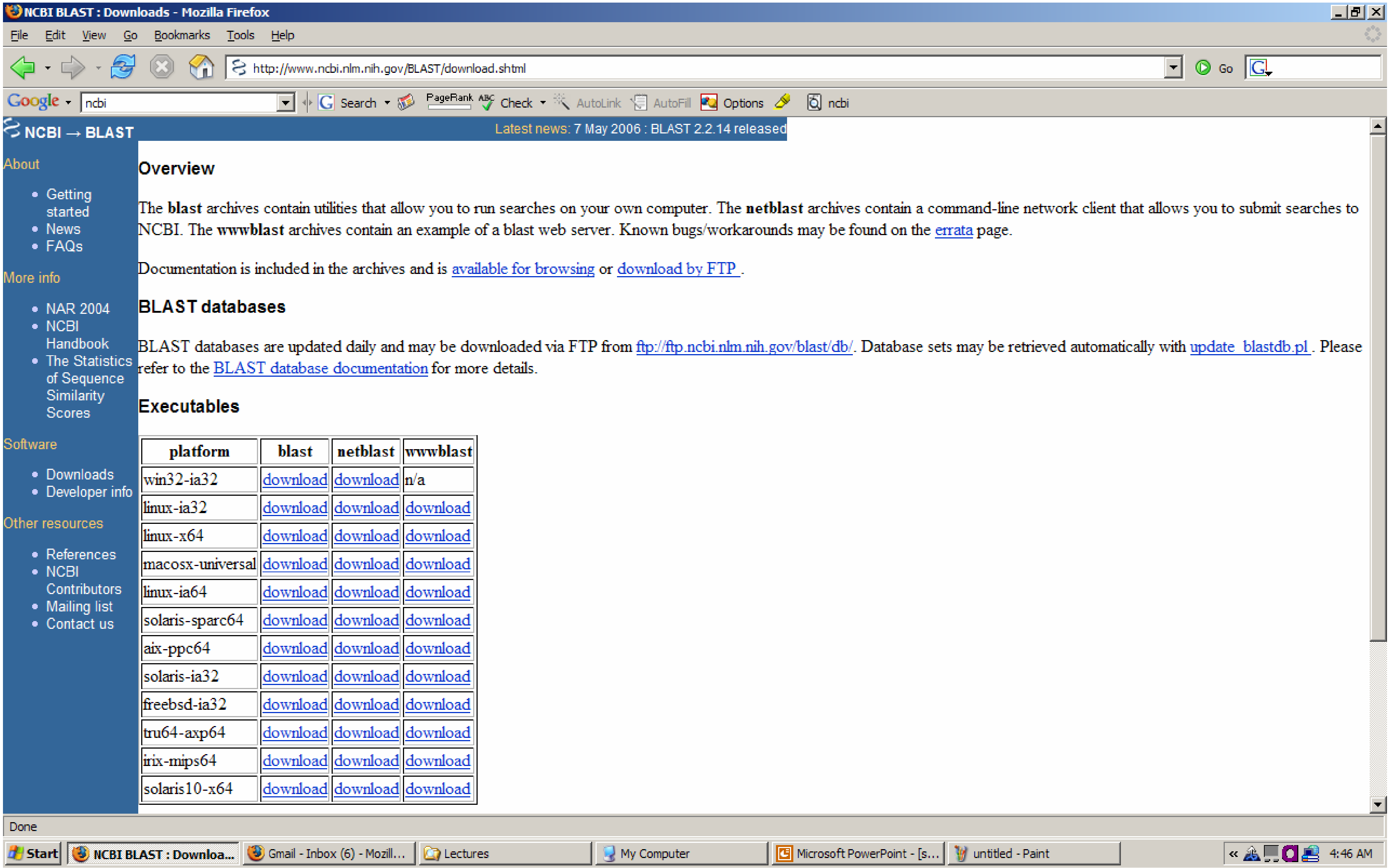
Matched sequences ranked by score and evalue



```
>□ gi|35213333|dbj|BAC90705.1| G gll2764 [Gloeobacter violaceus PCC 7421]
 gi|37522333|ref|NP_925710.1| G hypothetical protein gll2764 [Gloeobacter violaceus PCC 7421]
Length=559

 Score =  103 bits (256),  Expect = 5e-21, Method: Composition-based stats.
 Identities = 89/194 (45%), Positives = 115/194 (59%), Gaps = 12/194 (6%)

Query  7    KIMPVGDSCTEGMGGGEMGSYRTELYRLLTQAGLSIDFVGSQRSGPSSLPDKDHEGHSGW  66
            K+MP+GDS TEG      G YRT+L+  L    G + DFVGSQ SGPSSL DK+HEGH G+
Sbjct  108  KVMPLGDSITEGFTVS--GGYRTDLWNSLVSEGSNADFVGSQSSGPSSLSDKNHEGHPGY  165

Query  67   TIPQIASNINNWLNTHNPDVVFlwiggndlllngn--lnatglsnlIDQIFTVKPNVTLF  124
             I QIA  I++WL  + P+ V L IG ND+  N +        LS LIDQIF ++ +V L+
Sbjct  166  FIDQIADGIDDWLPKYKPETVLLLIGTNDIEKNNDPGGAPGRLSALIDQIFALRSSVKLY  225

Query  125  VADYYPWPE-AIKQ----YNAVIPGIVQQKANAGKKVYFVKLSEIQFDRNTDISWDGLHL  179
            VA   P  + AI Q    YNA IPGIV  K   GKKV +V +           D++ D +H
Sbjct  226  VASIPPADDSAINQRVLDYNAAIPGIVNGKITQGKKVVYVDIYNAL--TTADLA-DTVHP  282

Query  180  SEIGYKKIANIWYK  193
               GY KIA+ W++
Sbjct  283  DAEGYAKIADRWFE  296
```

Significant local alignments

Software and database download



Overview

The **blast** archives contain utilities that allow you to run searches on your own computer. The **netblast** archives contain a command-line network client that allows you to submit searches to NCBI. The **wwwblast** archives contain an example of a blast web server. Known bugs/workarounds may be found on the errata page.

Documentation is included in the archives and is available for browsing or download by FTP .

BLAST databases

BLAST databases are updated daily and may be downloaded via FTP from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. Database sets may be retrieved automatically with update_blastdb.pl . Please refer to the BLAST database documentation for more details.

Executables

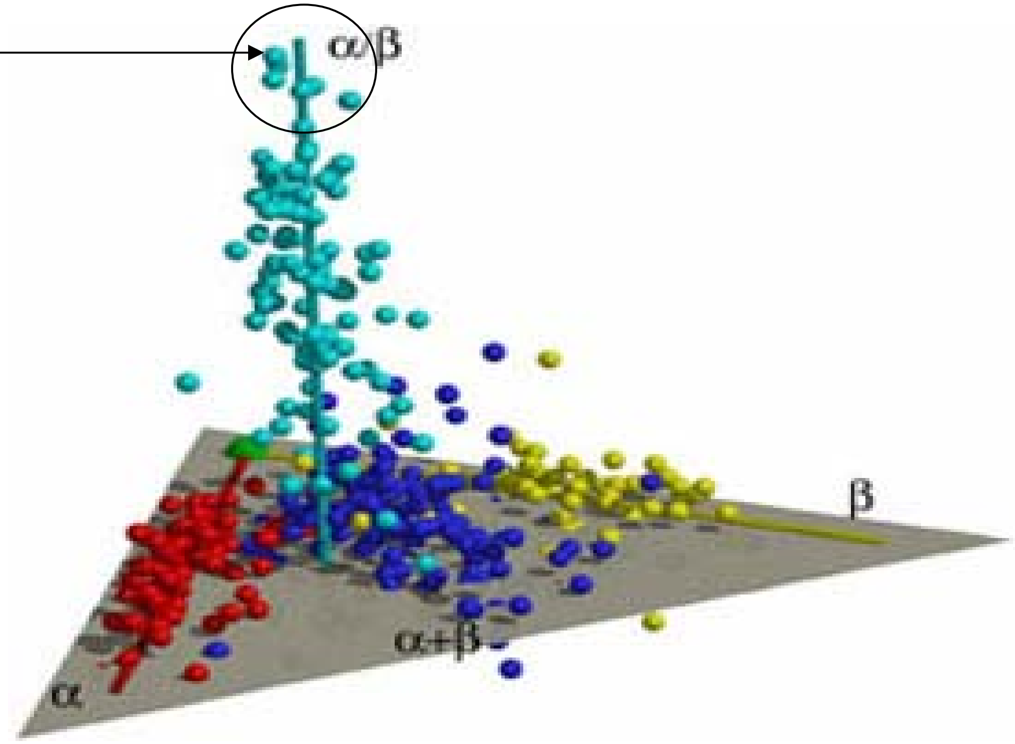
platform	blast	netblast	wwwblast
win32-ia32	download	download	n/a
linux-ia32	download	download	download
linux-x64	download	download	download
macosx-universal	download	download	download
linux-ia64	download	download	download
solaris-sparc64	download	download	download
aix-ppc64	download	download	download
solaris-ia32	download	download	download
freebsd-ia32	download	download	download
tru64-axp64	download	download	download
irix-mips64	download	download	download
solaris10-x64	download	download	download

Database Search Using Sequence Profiles

- Multiple Related Sequences (protein family and super family)
- Evolutionary relationship
- More data, more robust, more sensitive
- Consider a group of related sequences (profile) is a **POWERFUL** idea (sequence search, alignment, and protein structure prediction).

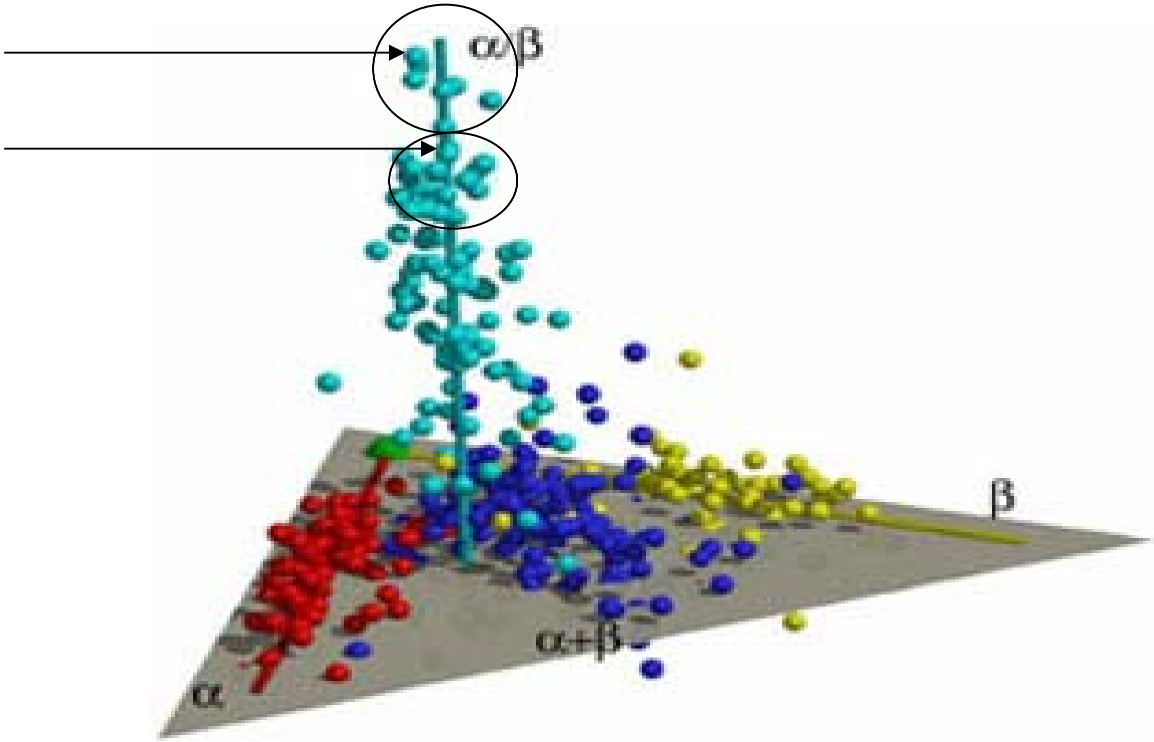
Why does a family of sequences help?

Step 1



Protein Universe

Why does a family of sequences help?

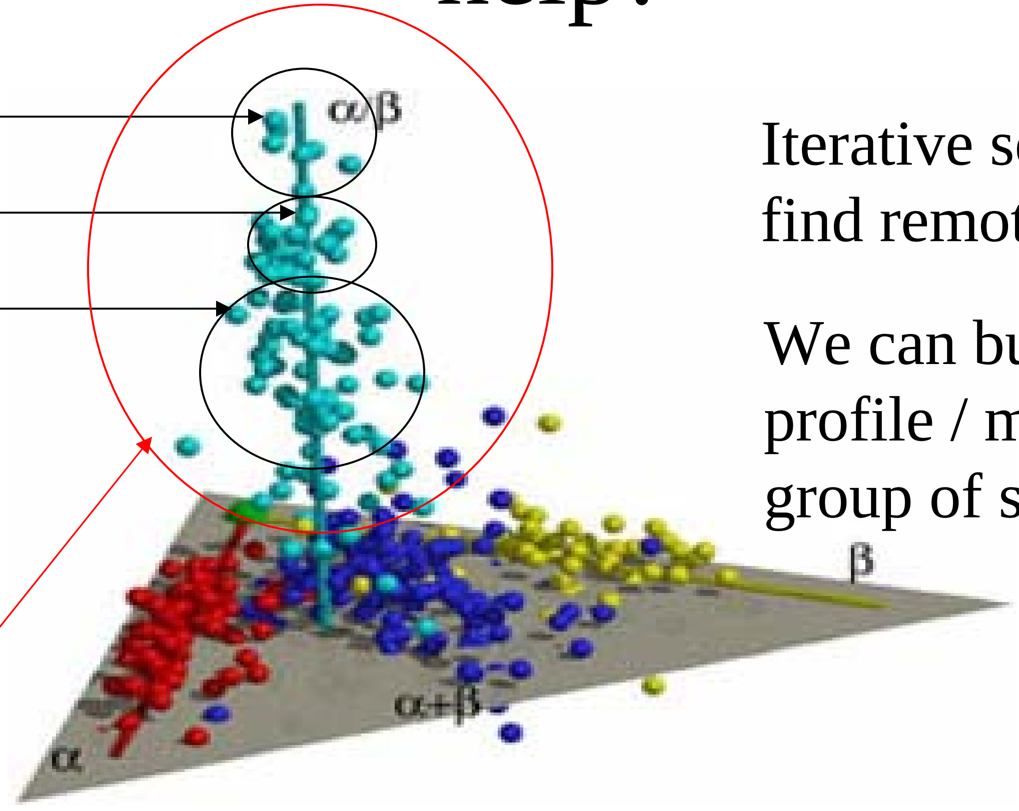


Step 1

Step 2

Protein Universe

Why does a family of sequences help?



Step 1

Step 2

Step 3

Protein Family

Protein Universe

Iterative search helps find remote homologs.

We can build a sophisticated profile / model from the group of sequences.

Representation of Profile

- Probability Matrix
- Hidden Markov Model
- Position Specific Scoring Matrix (PSSM)
- Profile can be considered a generalized sequence

Probability Matrix

Positions

123456789…

```
DSRSKGEPRDSGTLQSQEAKAVKKTSLFE
PRRKTVLSLFDEEEDKMEDQNIIQAPQKE
DSRSKGE-RDSGTLQSQEAKAVKKTSLFE
PRDKTVL-LFDEEEDKMEDQNIIQAPQKE
DSRSKGE-RD-GTLQSQEAKAVKKTSLFE
PRTKTVL-LF-EEEDKMEDQNIIQAPQKE
DSRSKGE-RD-GTLQSQEAKAVKKTSLFE
PRTKTVL-LFDEEEDKMEDQNIIQAPQKE
DSRSKGEPRDSGTLQSQEAKAVKKTSLFE
PSTKTVL-LFDEEEDKMEDQNIIQAPQKE
```

Multiple Sequence Alignment of a Protein Family

	A	C	D	E	…	P	Q	R	S	T	…
1	0	0	.5	0	0	.5	0	0	0	0	0
2	0	0	0	0	0	0	0	.4	.6	0	0
3											
4											
.											
.											
.											

Convert to L * 20 Probability Matrix

Profile captures more variations and conservations. It is more robust than single sequence. Essentially, it contains evolutionary information. Note: To avoid 0-probability, pseudo-count is used.

Hidden Markov Model

```
DSRSKGEPRDSGTLQSQEAKAVKKTSLFE
PRRKTVLSLFDEEEDKMEDQNIIQAPQKE
DSRSKGE-RDSGTLQSQEAKAVKKTSLFE
PRDKTVL-LFDEEEDKMEDQNIIQAPQKE
DSRSKGE-RD-GTLQSQEAKAVKKTSLFE
PRTKTVL-LF-EEEDKMEDQNIIQAPQKE
DSRSKGE-RD-GTLQSQEAKAVKKTSLFE
PRTKTVL-LFDEEEDKMEDQNIIQAPQKE
DSRSKGEPRDSGTLQSQEAKAVKKTSLFE
PSTKTVL-LFDEEEDKMEDQNIIQAPQKE
```

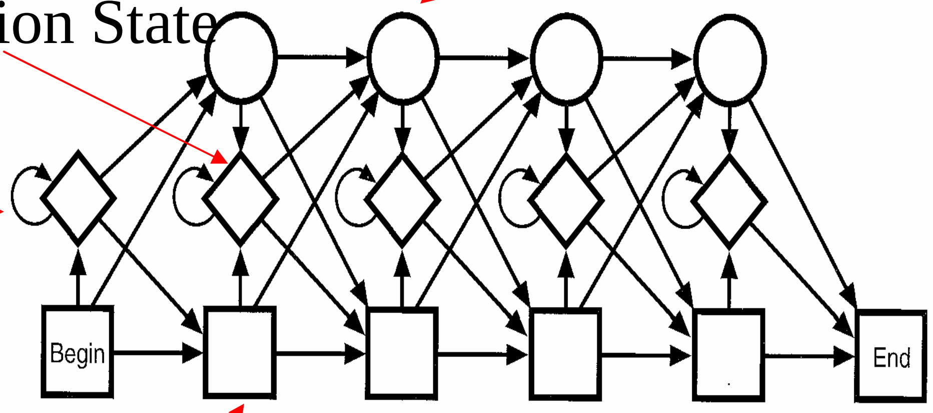
Deletion

Insertion

Conservation

Deletion State

Insertion State



Main State

Each main/insertion state has different probabilities of emitting amino acids.

Simplified Example of Position Specific Scoring Matrix

	A	C	D	E	…
1	0.1	0.05	0.5	0.01	
2					
3					
4					
.					
.					
.					

Background probability is 1/20 = 0.05

→

	A	C	D	E	…
1	0.30	0	1	-0.7	
2					
3					
4					
.					
.					
.					

Normalize observed probability (p_i) using background probability (q_i). Score = log (p_i / q_i). (log – odds)

PSI-BLAST Algorithm

- Use BLAST to search database. Use significantly matched sequences to construct a PSSM
- Repeat

  Use PSSM to search database

  Use significant matched sequences to construct a PSSM
- Until no new sequence is found or reach the maximum number of iterations.

Comments

- The algorithm of compare PSSM against sequence is the same as BLAST except that score is directly taken from PSSM instead of substitution matrix such as PAM or BLOSUM
- Sensitivity of PSI-BLAST is significantly improved over BLAST and other sequence only approaches such as Smith-Waterman sequence alignment method.

Use Standalone PSI-BLAST Software

- Download: http://130.14.29.110/BLAST/download.shtml
- Command:

```
blastpgp –i seq_file  –j iteration –h include_evalue_threshold –e report_evalue_threshold –d database –o output_file
```

-i: input sequence file in FASTA format
-j: number of iterations
-d: pre-formatted sequence database
-h: cut-off e-value of including a sequence into PSSM during iterations
-e: cut-off e-value of reporting a sequence
-o: output file

- Database

Use pre-formatted database such as NR (non-redundant protein sequence database) or your own database.
Format your own database using command: `formatdb –i sequence_file –o [T/F]`
T: created index using sequence id (can potentially speed up search).

Ten Topics

- 1. Introduction to Molecular Biology and Bioinformatics
- 2. Pairwise Sequence Alignment Using Dynamic Programming
- 3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
- 4. Multiple Sequence Alignment
- 5. Gene and Motif Identification
- 6. Phylogenetic Analysis
- 7. Protein Structure Analysis and Prediction
- 8. RNA Secondary Structure Prediction
- 9. Clustering and Classification of Gene Expression Data
- 10. Search and Mining of Biological Databases, Databanks, and Literature